Charotar University of Science and Technology (CHARUSAT)
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Department of Computer Science and Engineering
Lesson Planning Document (LPD)

1. GENERAL DETAILS

Faculty Name:	Priyanka Padhiyar	**Faculty Email:**	priyankapadhiyar.dcs@charusat.ac.in	**Department:**	Computer Science and Engineering (CSE)
Subject Code:	OCAIML4001	**Subject Name:**	RESPONSIBLE & SAFE AI SYSTEM	**Term Duration:**	23-06-2025 to 10-10-2025
Semester:	7th semester	**Division:**	Division 1 & Division 2	**Academic Year:**	2025-26
Lecture Hours/week:	3	**Lab Hours/week:**	2	**Credits:**	4
Course Prerequisites:	Fundamentals of Artificial Intelligence Machine Learning Basics Deep Learning Techniques NLP and Computer Vision Foundations Programming Skills Ethical and Legal Awareness (Basic) Mathematical Foundations				
Course Prerequisites Materials:	https://onlinecourses.nptel.ac.in/noc25_cs118/preview https://onlinecourses.nptel.ac.in/noc24_cs132/preview https://nptel.ac.in/courses/106106472				

2. UNIT DETAILS

Unit 1

Unit Name:	Introduction to Responsible AI & Emerging Risks	**Faculty Name:**	Priyanka Padhiyar
Start Date:	23/06/2025	**End Date:**	11/07/2025
No. of Lectures:	8	**CO Mapping:**	CO1, CO2, CO3, CO4
Unit Topics:	AI Capabilities Improvement in last 5-10 years , Imminent risks from AI Models: Toxicity, bias, goal misspecification, adversarial examples etc. , Long-term risks from AI Models: Misuse, Misgeneralization, Rogue AGI , Principles of RAI - Transparency; Accountability; Safety, Robustness and Reliability;Privacy and Security; Fairness and non-discrimination; Human-Centred Values; Inclusive and Sustainable development, Interpretability, Recap of Deep Learning Techniques, Language/Vision Models , AI Risks for Gen models, Adversarial Attacks – Vision, NLP, Superhuman Go agents.		
Self Study Topics:	N/A		
Self Study Materials:	N/A		
Teaching Pedagogy:	ICT based learning, Chalk and Talk, Demonstration/Simulation-Based Learning		
Skill Mapping:	Technical Skills		
Unit Materials:	https://onlinecourses.nptel.ac.in/noc25_cs118/preview https://onlinecourses.nptel.ac.in/noc24_cs132/preview https://nptel.ac.in/courses/106106472		
Skill Objectives:	Understand evolution and risks in AI systems; Apply principles of Responsible AI to real-world examples.		
Topics Beyond Unit:	RAI concepts like fairness, accountability, sustainability are part of AI policy and ethics—emerging topics.		
Interlink Topics:	N/A		

Unit 2

Unit Name:	AI Safety through Explainability & Transparency	**Faculty Name:**	Priyanka Padhiyar
Start Date:	14/07/2025	**End Date:**	29/07/2025
No. of Lectures:	8	**CO Mapping:**	CO1, CO3, CO2
Unit Topics:	ML Poisoning Attacks like Trojans ,Implications for current and future AI safety , Explainability ,Imminent and Long-term potential for transparency techniques, Mechanistic Interpretability , Representation Engineering, model editing and probing, Critiques of Transparency for AI Safety		
Self Study Topics:	Neuro-Symbolic Explainability		
Self Study Materials:	N/A		
Teaching Pedagogy:	ICT based learning, Demonstration/Simulation-Based Learning, Blended Learning		
Skill Mapping:	Technical Skills, Professional Skills		
Unit Materials:	https://onlinecourses.nptel.ac.in/noc25_cs118/preview https://onlinecourses.nptel.ac.in/noc24_cs132/preview https://nptel.ac.in/courses/106106472		
Skill Objectives:	Analyze vulnerabilities in ML models; Evaluate explainability approaches for AI safety.		
Topics Beyond Unit:	Real-World Case Studies of Trojan Attacks in AI Systems Counterfactual Explanations in AI Decision-Making Mechanistic Interpretability: Circuit-Level Analysis		
Interlink Topics:	N/A		

Unit 3

Unit Name:	Privacy and Fairness in AI Systems	**Faculty Name:**	Priyanka Padhiyar
Start Date:	30/07/2025	**End Date:**	05/08/2025
No. of Lectures:	7	**CO Mapping:**	CO2, CO3, CO4
Unit Topics:	Privacy & Fairness in AI		
Self Study Topics:	N/A		
Self Study Materials:	N/A		
Teaching Pedagogy:	ICT based learning, Chalk and Talk, Active Learning, Demonstration/Simulation-Based Learning, Experiential Learning		
Skill Mapping:	Technical Skills		
Unit Materials:	https://onlinecourses.nptel.ac.in/noc25_cs118/preview https://onlinecourses.nptel.ac.in/noc24_cs132/preview https://nptel.ac.in/courses/106106472		
Skill Objectives:	Apply fairness metrics; Interpret privacy concerns and propose mitigation strategies.		
Topics Beyond Unit:	Federated Learning and Privacy-Preserving AI		
Interlink Topics:	N/A		

Unit 4

Unit Name:	RAI Tools, Regulations, and Ethical Concerns	**Faculty Name:**	Priyanka Padhiyar
Start Date:	06/08/2025	**End Date:**	23/08/2025
No. of Lectures:	9	**CO Mapping:**	CO3, CO4, CO5, CO2
Unit Topics:	Metrics and Tools for RAI - measuring bias/fairness, adversarial testing, explanations (Lime/SHAP/GradCam), audit mechanisms ,Regulation landscape - DPDP act (India), GDPR (EU), EU AI act, US presidential declaration, Ethical approvals, informed consent, participatory design, future of work, Indian context ,What is AGI? When could it be achieved? ,Instrumental Convergence: Power Seeking, Deception etc.		
Self Study Topics:	N/A		
Self Study Materials:	N/A		
Teaching Pedagogy:	ICT based learning, Demonstration/Simulation-Based Learning, Active Learning		
Skill Mapping:	Technical Skills, Professional Skills		
Unit Materials:	https://onlinecourses.nptel.ac.in/noc25_cs118/preview https://onlinecourses.nptel.ac.in/noc24_cs132/preview https://nptel.ac.in/courses/106106472		
Skill Objectives:	Use tools for bias/explainability; Analyze ethical/regulatory frameworks affecting AI.		
Topics Beyond Unit:	Instrumental Convergence: Power-Seeking, Deception in AGI		
Interlink Topics:	N/A		

Unit 5

Unit Name:	Applications and Policy Aspects of RAI	**Faculty Name:**	Priyanka Padhiyar
Start Date:	25/08/2025	**End Date:**	05/09/2025
No. of Lectures:	6	**CO Mapping:**	CO2, CO3, CO4, CO5, CO6
Unit Topics:	RAI in Legal domain , RAI in Health care domain , RAI in Education domain , A few other domains ,Policy issues in RAI .		
Self Study Topics:	N/A		
Self Study Materials:	N/A		
Teaching Pedagogy:	ICT based learning, Demonstration/Simulation-Based Learning, Experiential Learning		
Skill Mapping:	Technical Skills		
Unit Materials:	https://onlinecourses.nptel.ac.in/noc25_cs118/preview https://onlinecourses.nptel.ac.in/noc24_cs132/preview https://nptel.ac.in/courses/106106472		
Skill Objectives:	Identify RAI challenges across sectors; Develop policy-aligned AI recommendations.		
Topics Beyond Unit:	RAI for Disaster Management and Climate Monitoring		
Interlink Topics:	N/A		

Unit 6

Unit Name:	Industry Perspectives and Future Directions	**Faculty Name:**	Priyanka Padhiyar
Start Date:	08/09/2025	**End Date:**	19/09/2025
No. of Lectures:	6	**CO Mapping:**	CO4, CO3, CO5, CO6, CO7
Unit Topics:	Couple of panel discussion with industry practitioners, academic, government (possibly), and others , Fireside chat with eminent personalities , Recorded Paper reading discussion		
Self Study Topics:	N/A		
Self Study Materials:	N/A		
Teaching Pedagogy:	ICT based learning, Demonstration/Simulation-Based Learning, Active Learning, Experiential Learning		
Skill Mapping:	Technical Skills		
Unit Materials:	https://onlinecourses.nptel.ac.in/noc25_cs118/preview https://onlinecourses.nptel.ac.in/noc24_cs132/preview https://nptel.ac.in/courses/106106472		
Skill Objectives:	Engage with expert perspectives; Develop awareness of current RAI trends through scholarly discussions.		
Topics Beyond Unit:	The Future of AI Governance in India: From Ethics to Enforcement.		
Interlink Topics:	N/A		

3. PRACTICAL DETAILS

Practical 1

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	5
Probable Week:	Week 1 (23-06-2025 - 29-06-2025)	**CO Mapping:**	CO1
Practical Aim:	Uncovering Bias in Social Media Sentiment Analysis An AI ethics researcher at a startup is tasked with auditing a sentiment analysis tool developed for a major social media platform. This tool is designed to automatically filter toxic comments and flag negative posts. However, complaints have arisen indicating that the system disproportionately flags content related to certain identities (e.g., comments containing gender or race-specific terms), even when such content is not offensive. The objective is to identify, analyze, and report any bias present in the system and propose improvements using fairness metrics and natural language processing techniques.		
Practical Tasks:	Collect sample social media comments (or use pre-annotated datasets such as Twitter Hate Speech, Toxic Comments dataset, etc.). Apply the existing sentiment analysis model (open-source like VADER, TextBlob, or a mock ML model). Analyze model behavior on: Identity-specific comments (e.g., involving words like “woman”, “Black”, “gay”) Neutral vs. flagged results Use fairness metrics such as: False Positive Rate (FPR) per identity group Disparate Impact Ratio Demographic Parity Visualize patterns of bias using confusion matrix or bar graphs. Propose improvements: Pre-processing techniques Balanced training data Bias-aware models like Fairlearn, AIF360		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	Introduction to Responsible AI & Emerging Risks		
Blooms Taxonomy:	Apply, Create		
Skill Mapping:	Technical Skills		
Skill Objectives:	Use sentiment analysis tools to identify biased predictions Apply fairness metrics to assess disproportionate impact Conduct critical evaluation of algorithmic decision-making Propose and test bias-reduction strategies		
Reference Material:	https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data		
Software/Hardware Requirements:	Anaconda		

Practical 2

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	5
Probable Week:	Week 3 (07-07-2025 - 13-07-2025)	**CO Mapping:**	CO2
Practical Aim:	Preventing Toxic Outputs in AI-Powered Language Systems. A social media company is preparing to launch an AI-powered virtual assistant that helps users compose posts and comments. During internal testing, the assistant occasionally generates offensive, abusive, or politically sensitive content in response to seemingly neutral or provocative prompts. Concerned about user safety and brand reputation, the company assigns the AI Ethics and Safety Team to audit the model's output, detect toxic behavior, and recommend mitigation strategies before deployment.		
Practical Tasks:	Analyze outputs of a language model (e.g., GPT-2, open-source LLMs, or simulated outputs) in response to diverse prompts. Identify examples of toxic, biased, or politically sensitive content. Use tools like Perspective API or ToxiScore for automated toxicity detection. Manually annotate and classify outputs as acceptable, borderline, or toxic. Explore prompt engineering techniques or filtering methods to reduce toxic outputs. Propose ethical, design, and technical interventions for mitigating harm. Submit a brief audit report with findings and recommendations.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	AI Safety through Explainability & Transparency		
Blooms Taxonomy:	Apply, Create		
Skill Mapping:	Technical Skills		
Skill Objectives:	Identify and classify potentially harmful AI outputs Use automated tools to detect toxicity Design mitigation strategies using prompt filtering and output constraints Apply ethical principles to real-world AI systems Draft concise audit reports and reflect on AI safety		
Reference Material:	https://perspectiveapi.com/		
Software/Hardware Requirements:	Anaconda		

Practical 3

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	4
Probable Week:	Week 5 (21-07-2025 - 27-07-2025)	**CO Mapping:**	CO3
Practical Aim:	Defending Vision Models Against Adversarial Attacks in High Stakes Applications. CO3, CO4/ PO3, PO4 A healthcare startup is deploying an AI system for automated medical image diagnosis, using deep learning to identify diseases from X-rays and CT scans. During security evaluation, it is found that small, imperceptible changes to the images can trick the model into making incorrect diagnoses, such as missing a tumor or predicting disease in a healthy image. To ensure safety and reliability, the development team is assigned to perform robustness testing using adversarial attacks and implement defense mechanisms to improve the model's resilience.		
Practical Tasks:	A healthcare startup is deploying an AI system for diagnosing diseases using X-rays and CT scans. However, security evaluations reveal that minor pixel-level perturbations—imperceptible to humans—can lead to incorrect model predictions (e.g., misdiagnosis of tumors). The development team must: Evaluate the model’s vulnerability to adversarial attacks (e.g., FGSM, PGD). Generate adversarial samples using known attack techniques. Apply defense mechanisms like adversarial training, input preprocessing, or model regularization. Analyze the impact of defenses on model accuracy and robustness. Deliverables: Report with before/after results on attack success rate and model performance. Code demonstrating adversarial attack and defense implementation. Reflection on applicability in healthcare domain and ethical considerations.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	Privacy and Fairness in AI Systems		
Blooms Taxonomy:	Apply		
Skill Mapping:	Technical Skills		
Skill Objectives:	Understand and simulate adversarial attacks on vision models. Develop practical defense strategies to enhance model robustness. Analyze and interpret the results in the context of high-stakes domains like healthcare. Cultivate ethical awareness related to the deployment of AI systems.		
Reference Material:	https://paperswithcode.com/task/adversarial-defense		
Software/Hardware Requirements:	Anaconda		

Practical 4

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	3
Probable Week:	Week 7 (04-08-2025 - 10-08-2025)	**CO Mapping:**	CO4
Practical Aim:	Detect hidden trojans in neural networks using trigger inputs and specialized datasets to reveal security vulnerabilities. A defense technology company plans to deploy a pre-trained AI model for object detection in surveillance drones. This model was outsourced from an external vendor. During testing, unusual behavior is observed—certain inputs with specific patterns consistently trigger incorrect predictions, such as identifying a “stop sign” as a “green light.” Suspecting a backdoor (trojan) attack, the company assigns the AI Security Team to conduct a forensic audit of the model. The goal is to simulate and detect trojans using trigger-based inputs, compare behavior between clean and trojaned models, and assess potential security vulnerabilities before deployment.		
Practical Tasks:	A defense technology company is preparing to deploy a pre-trained object detection model for surveillance drones. The model, developed by a third-party vendor, exhibits suspicious behavior: certain patterned inputs cause it to misclassify objects (e.g., a stop sign misidentified as a green light). The AI Security Team is tasked with: Auditing the model for potential trojan (backdoor) attacks. Simulating known backdoor attacks on clean models using public datasets. Crafting trigger inputs and observing output variations. Comparing performance between clean and potentially trojaned models. Analyzing misclassification rates and identifying vulnerabilities. Documenting findings and suggesting mitigation strategies.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	RAI Tools, Regulations, and Ethical Concerns		
Blooms Taxonomy:	Apply		
Skill Mapping:	Technical Skills		
Skill Objectives:	Understand trojan (backdoor) attack mechanisms in AI models. Develop trigger-based inputs to expose vulnerabilities. Compare and analyze model predictions under normal and triggered conditions. Simulate a backdoor attack on a sample model. Propose mitigation strategies to improve model robustness.		
Reference Material:	https://towardsdatascience.com/backdoor-attacks-on-deep-learning-models-74dcf2d72a8b		
Software/Hardware Requirements:	Anaconda		

Practical 5

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	4
Probable Week:	Week 9 (18-08-2025 - 24-08-2025)	**CO Mapping:**	CO5
Practical Aim:	Model Explainability with LIME and SHAP on Text Classification A fintech company has developed a sentiment analysis system to evaluate customer feedback and automatically escalate negative reviews for priority handling. However, the customer support team raises concerns that some negative reviews are being misclassified or misunderstood by the model, leading to inconsistent escalation. To address this, the AI team is tasked with making the model’s decisions explainable using tools like LIME and SHAP. The goal is to interpret how input features (words/phrases) influence model predictions, compare local vs. global explanations, and identify cases of model bias or misclassification—thereby improving transparency and trustworthiness of the system.		
Practical Tasks:	A fintech company uses a sentiment analysis model to classify customer feedback as positive, neutral, or negative. However, some critical negative feedback is being misclassified, leading to customer dissatisfaction. The AI team must integrate LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) to: Explain model predictions for individual instances (local interpretability). Understand overall feature importance (global interpretability). Detect patterns of bias or error in misclassified examples. Improve the model or escalation logic using explainability insights. You will: Train or use a pre-trained sentiment classifier (e.g., logistic regression or LSTM). Use LIME and SHAP to visualize how specific words/phrases influenced predictions. Analyze differences in explanations for correct vs. incorrect predictions. Report how explainability helps improve decision transparency and error handling.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	RAI Tools, Regulations, and Ethical Concerns		
Blooms Taxonomy:	Apply		
Skill Mapping:	Technical Skills		
Skill Objectives:	Use LIME and SHAP for model interpretation in NLP context Explain differences between local and global interpretability Identify misleading or biased model behaviors Communicate technical findings to non-technical stakeholders (e.g., customer support) Recommend strategies to improve system transparency and trust		
Reference Material:	https://github.com/marcotcr/lime https://github.com/slundberg/shap		
Software/Hardware Requirements:	Anaconda		

Practical 6

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	2
Probable Week:	Week 9 (18-08-2025 - 24-08-2025)	**CO Mapping:**	CO6
Practical Aim:	Fair Classification and Bias Mitigation using Fairlearn Toolkit A government agency is deploying an AI system to automate screening of applicants for financial aid. During testing, the model shows disproportionately lower approval rates for certain racial and gender groups. To ensure fairness and regulatory compliance, the data science team is assigned to audit the model’s decisions and apply bias mitigation techniques using fairness toolkits.		
Practical Tasks:	Load and explore a dataset (e.g., financial aid, loan approval, or UCI Adult Income dataset). Train a baseline classification model (e.g., Logistic Regression, Random Forest) to predict approval outcomes. Analyze model performance and fairness metrics across protected attributes (e.g., race, gender). Use Fairlearn toolkit to: Audit the model using fairness metrics (e.g., Demographic Parity, Equal Opportunity). Apply mitigation strategies like Exponentiated Gradient Reduction or Grid Search Reduction. Compare fairness and accuracy trade-offs before and after mitigation. Visualize disparities and fairness improvements.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	Applications and Policy Aspects of RAI		
Blooms Taxonomy:	Apply		
Skill Mapping:	Technical Skills		
Skill Objectives:	Understand the importance of fairness in AI decision systems. Learn to identify and quantify algorithmic bias. Apply appropriate fairness metrics and mitigation techniques. Communicate bias mitigation strategies clearly and ethically.		
Reference Material:	https://fairlearn.org/ https://archive.ics.uci.edu/ml/datasets/adult		
Software/Hardware Requirements:	Anaconda		

Practical 7

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	1
Probable Week:	Week 10 (25-08-2025 - 31-08-2025)	**CO Mapping:**	CO7
Practical Aim:	Differential privacy limits the information that models can leak about individual training samples. A health-tech startup is building an AI model to predict the risk of chronic diseases based on sensitive patient data like age, medical history, and lifestyle. While the model performs well, the company must ensure it complies with strict data privacy regulations such as HIPAA and GDPR. The concern is that attackers could extract information about individual patients from the trained model. To address this, the engineering team is assigned to implement Differential Privacy (DP) using the Opacus library in PyTorch, evaluate how much privacy is gained (ε), and understand the impact of privacy constraints on model performance.		
Practical Tasks:	A health-tech startup is building an AI model to predict chronic disease risk using patient data (age, medical history, lifestyle, etc.). Although the model is accurate, there's a risk of leaking sensitive information. The objective is to integrate Differential Privacy (DP) using Opacus, a PyTorch library, to ensure compliance with HIPAA and GDPR. Tasks include: Building a baseline model (e.g., logistic regression or simple MLP) on synthetic patient data. Integrating Opacus to train the model with DP-SGD. Comparing model performance (accuracy, F1-score) with and without DP. Calculating and interpreting the privacy budget (ε, δ). Visualizing the tradeoff between privacy and model performance.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	Applications and Policy Aspects of RAI		
Blooms Taxonomy:	Apply		
Skill Mapping:	Technical Skills		
Skill Objectives:	To understand and apply Differential Privacy principles in ML. To utilize Opacus for private model training. To critically evaluate privacy-performance tradeoffs. To develop ethical AI systems conforming to legal regulations.		
Reference Material:	https://opacus.ai/docs/		
Software/Hardware Requirements:	Anaconda		

Practical 8

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	4
Probable Week:	Week 10 (25-08-2025 - 31-08-2025)	**CO Mapping:**	CO7
Practical Aim:	Designing a Transparent and Inclusive Mental Health Chatbot A healthcare startup is developing an AI-powered mental health assistant chatbot to support users experiencing symptoms of depression and anxiety. The assistant analyzes user conversations and offers emotional support and recommendations. However, due to the sensitive nature of mental health data, ethical risks such as lack of informed consent, misdiagnosis, data misuse, and exclusion of vulnerable groups must be addressed before deployment. To ensure responsible development, the startup's AI Ethics team is tasked with creating a prototype of the assistant that integrates informed consent, participatory feedback, and ethical safeguards through user-centered design		
Practical Tasks:	A healthcare startup is creating an AI-based chatbot to provide emotional support to users with symptoms of depression and anxiety. The goal is to: Ensure informed consent before interaction. Incorporate participatory design with feedback loops from diverse user groups. Avoid bias and misdiagnosis by ensuring explainable and non-prescriptive recommendations. Protect user privacy and mental health data. Provide culturally sensitive and inclusive support. Task: Design a low-fidelity or functional prototype of the chatbot with: Transparent conversational flow (consent, disclaimers). Ethical disclaimers and referral to professionals. Sample intents and safe response templates. Mechanisms for participatory feedback. Accessibility features and inclusion considerations.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	Applications and Policy Aspects of RAI		
Blooms Taxonomy:	Apply, Create		
Skill Mapping:	Technical Skills		
Skill Objectives:	Design ethically informed conversational AI. Implement consent and privacy frameworks in prototypes. Demonstrate inclusive and user-centered chatbot features. Apply participatory methods to refine AI design. Explain ethical choices and safety considerations effectively.		
Reference Material:	UNESCO AI Ethics Guidelines (2021) The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems "Ethically Aligned Design" by IEEE "Responsible AI in Practice" – NPTEL Course by Prof. Ponnurangam Kumaraguru AI Now Institute Reports "Design Justice: Community-Led Practices" – Sasha Costanza-Chock		
Software/Hardware Requirements:	Anaconda		

Practical 9

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	4
Probable Week:	Week 11 (01-09-2025 - 07-09-2025)	**CO Mapping:**	CO6, CO7
Practical Aim:	Auditing Bias in Algorithmic Risk Prediction Tools: The COMPAS Controversy A state-level criminal justice department is using the COMPAS algorithm to assess the likelihood of reoffending among arrested individuals. The goal is to support judges in making bail and sentencing decisions. However, investigations by independent journalists and researchers have revealed that the system disproportionately assigns higher risk scores to Black defendants compared to White defendants, raising concerns of algorithmic bias and structural discrimination.To prepare for a public ethics review, the department commissions a data science team to analyze the fairness of the COMPAS system, propose mitigation strategies, and simulate a panel discussion on the ethical implications of deploying AI in the legal system.		
Practical Tasks:	A state-level criminal justice department uses the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm to predict the likelihood of a defendant reoffending. Reports indicate racial bias in the algorithm, particularly that Black defendants are more frequently labeled as high-risk than White defendants. Your team, acting as ethical auditors and data scientists, must: Analyze COMPAS dataset for evidence of bias using fairness metrics. Simulate a panel discussion on ethical implications and legal ramifications. Propose bias mitigation techniques to improve the algorithm’s fairness. Reflect on whether such systems should be used in judicial decisions.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	Industry Perspectives and Future Directions		
Blooms Taxonomy:	Apply		
Skill Mapping:	Technical Skills		
Skill Objectives:	Develop ability to audit algorithmic fairness using statistical tools. Understand and communicate ethical risks in AI-powered legal systems. Simulate multi-stakeholder discussions for AI governance. Propose actionable bias mitigation techniques.		
Reference Material:	ProPublica article: “Machine Bias” – https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems YouTube: COMPAS Algorithm Explained – [Various panel discussions]		
Software/Hardware Requirements:	Anaconda		

Practical 10

Faculty Name:	Priyanka Padhiyar	**Lab Hours:**	3
Probable Week:	Week 12 (08-09-2025 - 14-09-2025)	**CO Mapping:**	CO7
Practical Aim:	Interpreting Deep Learning Decisions in Critical AI Systems Using GradCAM A medical AI company is deploying a deep learning model to detect pneumonia in chest X-rays. The model shows high accuracy, but doctors are reluctant to trust it because the predictions lack transparency. During validation, some misclassifications reveal that the model is focusing on irrelevant features like hospital tags, corners, or background noise instead of lung regions. To improve trust, transparency, and ethical safety, the internal AI audit team is assigned to apply GradCAM on the trained model to visualize its focus areas, assess alignment with domain knowledge, and recommend improvements for more responsible use of the model.		
Practical Tasks:	A deep learning model trained to detect pneumonia from chest X-rays has shown high performance metrics, but doctors question its reliability due to opaque decision-making. Some misclassifications during validation indicate the model may be learning from spurious features such as hospital tags or image corners instead of lung regions. Task: Apply GradCAM (Gradient-weighted Class Activation Mapping) to visualize the attention regions of the model on chest X-ray images. Interpret whether the model is focusing on relevant anatomical features (lungs) or irrelevant regions. Based on observations, suggest improvements for more transparent and ethical AI usage in the medical field.		
Practical Pedagogy:	Problem-Based/Case Study Learning		
Evaluation Methods:	Viva, Code Review, Lab Performance, File Submission		
Associated Units:	Industry Perspectives and Future Directions		
Blooms Taxonomy:	Apply, Create		
Skill Mapping:	Technical Skills		
Skill Objectives:	Understand the concept and implementation of GradCAM. Interpret model decision-making regions using visual tools. Detect spurious correlations learned by models in real-world datasets. Communicate technical findings with ethical and domain relevance. Recommend mitigation strategies to increase trust and reliability in critical AI systems.		
Reference Material:	Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", ICCV 2017. Stanford CheXNet Pneumonia Detection paper and dataset. Articles on Explainable AI in Healthcare. Responsible AI Guidelines by WHO / IEEE. Tools: GradCAM Tutorial - PyTorch		
Software/Hardware Requirements:	Anaconda		

4. CIE DETAILS

No.	Unit Covered	Date	Marks	Duration (mins)	Evaluation Type	Bloom's Taxonomy	Evaluation Pedagogy	CO/PSO/PEO	Skills
1	-	02/07/2025	20 Marks	60 mins	Course Prerequisites CIE	Analyze, Apply	Problem-Based Evaluation	CO1, CO2, CO3, CO4, CO5, CO6, CO7, PSO1	Technical Skills
2	1	28/07/2025	20 Marks	60 mins	Lecture CIE	Evaluate, Apply	Short/Descriptive Evaluation	CO1, CO2	Communication Skills
3	2	05/09/2025	20 Marks	60 mins	Lecture CIE	Apply, Analyze, Evaluate	Group/Team Assessment	CO1	Creativity and Design Thinking Skills
		Total	**40 Marks**	**2 hours**					
4	-	03/10/2025	20 Marks	60 mins	Internal Practical	Analyze, Apply, Evaluate, Create, Understand	Problem-Based Evaluation	CO1, CO2, CO3, CO4, CO5, CO6, CO7	Technical Skills
5	-	10/10/2025	20 Marks	60 mins	Practical CIE	Analyze, Evaluate, Apply, Create	Problem-Based Evaluation	CO1, CO2, CO3	Technical Skills
		Total	**40 Marks**	**2 hours**					
6	1, 2, 3	06/10/2025	20 Marks	60 mins	Mid-term/Internal Exam	Analyze, Apply	Oral/Visual Communication-Based Evaluation (Presentation/Public Speaking/Viva)	CO1, CO2, CO3	Communication Skills

Overall Total	**120 Marks**	**6 hours**	

5. ADDITIONAL DETAILS

Academic Integrity:	Plagiarism, fabrication of reflections, or unethical use of AI tools will result in disciplinary action. Students must cite all tools, datasets, or models used. Group collaboration is allowed only when explicitly stated. Submitting identical work will result in a zero.
Attendance Policy:	Minimum 75% attendance is required to qualify for evaluations.
CIE Guidelines:	Out of 5 CIEs Best 4 CIEs will be consider
Classroom Conduct:	Maintain a respectful, inclusive, and collaborative environment. Laptops must be brought for every lab session. Active participation in discussions, ethical case studies, and group exercises is expected. Use of AI tools (e.g., ChatGPT, Copilot) must be disclosed if used in assignments or discussions. Maintain professional behavior when discussing socially sensitive topics.
Communication Channels:	Google Classroom: Material, announcements, deadlines Email: One-on-one communication

This LPD was downloaded on 17/07/2025 by Priyanka Padhiyar